*Biometrika* (2017), **xx**, x, *pp*. 1–8
*Printed in Great Britain*

# Pseudo-marginal Metropolis–Hastings using averages of unbiased estimators

BY CHRIS SHERLOCK
*Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K.*
c.sherlock@lancaster.ac.uk

ALEXANDRE H. THIERY
*Department of Statistics and Applied Probability, National University of Singapore, Singapore 117543.*
a.h.thiery@nus.edu.sg

AND ANTHONY LEE
*Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.*
anthony.lee@warwick.ac.uk

SUMMARY

We consider a pseudo-marginal Metropolis–Hastings kernel $\mathsf{P}_m$ that is constructed using an average of $m$ exchangeable random variables, as well as an analogous kernel $\mathsf{P}_s$ that averages $s < m$ of these same random variables. Using an embedding technique to facilitate comparisons, we show that the asymptotic variances of ergodic averages associated with $\mathsf{P}_m$ are lower bounded in terms of those associated with $\mathsf{P}_s$. We show that the bound provided is tight and disprove a conjecture that when the random variables to be averaged are independent, the asymptotic variance under $\mathsf{P}_m$ is never less than $s/m$ times the variance under $\mathsf{P}_s$. The conjecture does, however, hold when considering continuous-time Markov chains. These results imply that if the computational cost of the algorithm is proportional to $m$, it is often better to set $m = 1$. We provide intuition as to why these findings differ so markedly from recent results for pseudo-marginal kernels employing particle filter approximations. Our results are exemplified through two simulation studies; in the first the computational cost is effectively proportional to $m$ and in the second there is a considerable start-up cost at each iteration.

*Some key words*: Markov Chain Monte Carlo; Pseudo Marginal Markov Chain Monte Carlo; Importance Sampling

## 1. INTRODUCTION

The Metropolis–Hastings algorithm is often used to approximate expectations with respect to posterior distributions, making use of point-wise evaluations of the posterior density $\pi$ up to an arbitrary constant of proportionality. In cases where such evaluations are infeasible, the pseudo-marginal Metropolis–Hastings algorithm (Beaumont, 2003; Andrieu & Roberts, 2009) can be used if a realisation of a non-negative, unbiased stochastic estimator of the target density, possibly up to an unknown normalisation constant, is available. These estimators can, for example, be constructed using importance sampling (Beaumont, 2003), a particle filter or sequential Monte Carlo (Andrieu et al., 2010).

A key tuning parameter of such pseudo-marginal algorithms is the number of samples or particles, which we denote by $m$, and we are interested in the relationship between $m$ and the computational efficiency of the pseudo-marginal algorithm for approximating posterior expectations. The algorithm is a type of Markov chain Monte Carlo method: a Markov chain with stationary distribution $\pi$ is simulated for a finite number of steps in order to compute an appropriately normalized partial sum. This quantity then serves as an approximation of a limiting ergodic average that is almost surely equal to the expectation of interest. One natural measure of computational inefficiency, defined precisely in the sequel, is the asymptotic variance of the ergodic average of interest multiplied by the computational effort required to simulate each value of the Markov chain. Andrieu & Vihola (2015) showed that the asymptotic variance for a pseudo-marginal algorithm is bounded below by the asymptotic variance of an idealized algorithm in which $\pi$ is evaluated exactly, so one can think of the relative asymptotic variance of a pseudo-marginal ergodic average as its asymptotic variance divided by the asymptotic variance of the idealized ergodic average. In some sense, this idealized algorithm is approached as $m \uparrow \infty$ and one might suppose that the relative asymptotic variance should therefore decrease to 1 as $m$ increases. This is indeed true for estimators that arise from importance sampling (Andrieu & Vihola, 2016), at least when the asymptotic variance is finite for some finite $m$. An important issue, then, is whether the decrease in asymptotic variance at the expense of increased computational effort is justified in terms of computational efficiency.

In this article, we consider arbitrary pseudo-marginal algorithms where the posterior density is estimated using an average of unbiased estimators, such as with importance sampling, and we show that in all such cases the asymptotic variance when $m$ samples are used is not much smaller than the asymptotic variance when a single sample is used, divided by $m$. Thus if the computational cost is roughly proportional to $m$, as is often the case, there is little, if any, gain in using more than one sample. This formalises empirical observations made in Section 3.4 of an early version of Sherlock et al. (2015) (arXiv reference 1309.7209v1), and generalizes the main result of Bornn et al. (2017), which assumes that the pseudo-marginal kernels are positive and the estimators to be averaged are independent and take only one non-zero value. We demonstrate that our bound is tight and illustrate it through two simulation studies. In the second study, an additional fixed and large cost is associated with simulating $m$ samples, so that the computational efficiency is maximized at some $m \gg 1$. Our result also demonstrates that asymptotic variance being infinite for $m = 1$ implies that it is infinite for all finite $m$. The theory also suggests that when $m_0$ estimates can be obtained in parallel at no additional cost, e.g. by using vectorized instructions, multiple processor cores or distributed computing, it should be close to optimal to obtain $m_0$ estimates and then average these. Our result and the concomitant advice, which apply to a single average, differ markedly from the results and advice for particle filters, where the unbiased estimator is a product of averages.

We adopt the notation $x \wedge y = \min\{x, y\}$; for an integer $m \geq 0$, we set $[m] \equiv \{1, 2, \ldots, m\}$. For a probability measure $\pi$ on some measurable space $(\mathsf{X}, \Sigma)$ and a $\pi$-integrable test function $\varphi : \mathsf{X} \to \mathbb{R}$, we define $\pi(\varphi) \equiv \int_{\mathsf{X}} \varphi(x)\, \pi(\mathrm{d}x)$ and use the notation $L^2(\pi) \equiv \left\{\varphi : \mathsf{X} \to \mathbb{R} \,:\, \pi(\varphi^2) < \infty\right\}$ to designate the usual Hilbert space with norm $\|\varphi\|_\pi^2 = \pi(\varphi^2)$.

## 2. MAIN RESULTS

### 2·1. *Asymptotic variance of ergodic averages and accept-reject kernels*

Consider a Markov transition kernel $P$ with invariant distribution $\pi$ and associated Markov chain $\{X_k\}_{k=0}^{\infty}$ with $X_0 \sim \pi$. For any $\varphi \in L^2(\pi)$, an estimator of $\pi(\varphi)$ is the ergodic average

$n^{-1}\sum_{k=1}^{n}\varphi(X_k)$ and the asymptotic variance of the ergodic average is

$$\mathrm{V}(\varphi, P) \equiv \lim_{n\to\infty} \mathrm{var}\left\{ n^{-1/2} \sum_{k=1}^{n} \varphi(X_k) \right\}.$$

All Markov kernels $P$ in this article involve accepting or rejecting a sample from a proposal kernel $Q$ according to an acceptance probability $\alpha(x; x')$. We define the marginal acceptance probability from $x$, $\alpha(x) = \int_{\mathsf{X}} \alpha(x; x')\, Q(x, \mathrm{d}x')$. The kernel $P$ is then of the form

$$P(x, \mathrm{d}x') \equiv \{1 - \alpha(x)\}\, \delta_x(\mathrm{d}x') + Q(x, \mathrm{d}x')\, \alpha(x; x'). \tag{1}$$

### 2·2. *Pseudo-marginal Metropolis–Hastings*

Let $\pi(\mathrm{d}x) = \pi(x)\,\mathrm{d}x$ be a probability distribution on $\mathsf{X}$, where $\mathrm{d}x$ denotes a dominating measure, and $Q$ be a proposal kernel with density $q$, i.e. $Q(x, \mathrm{d}x') = q(x, x')\,\mathrm{d}x'$. The $\pi$-reversible Metropolis–Hastings kernel associated with $Q$ is defined via (1) by taking the acceptance probability $\alpha(x; x') \equiv 1 \wedge r(x, x')$ where $r(x, x')$ is the Metropolis–Hastings ratio,

$$r(x, x') \equiv \frac{\pi(x')\, q(x', x)}{\pi(x)\, q(x, x')}.$$

When pointwise relative evaluation of the density $\pi$ is not possible, $\alpha(x; x')$ is intractable. The pseudo-marginal Metropolis–Hastings algorithm introduces an unbiased approximation to the posterior, $\widehat{\pi}(x, U)$, where $U$ is a vector of auxiliary random variables. We will be interested in the random variable $W \equiv \widehat{\pi}(x, U)/\pi(x) \in \mathsf{W} \subseteq [0, \infty)$ which satisfies $\mathrm{E}(W) = 1$. The pseudo-marginal algorithm simulates a Metropolis–Hastings Markov chain on the extended state space $\mathsf{X} \times \mathsf{W}$ with proposal density $q(x, x')\, q_{x'}(w')$ and invariant density $\widetilde{\pi}(x, w) = \pi(x) q_x(w) w$, i.e. proposals are accepted with the usual Metropolis–Hastings ratio, which in this case reads $\alpha(x, w; x', w') := 1 \wedge \{r(x, x')\, w'/w\}$. Importantly, $\widetilde{\pi}$ admits $\pi$ as its $x$-marginal.

### 2·3. *Pseudo-marginal algorithm using averages*

Suppose that for each $x \in \mathsf{X}$ it is possible to generate an unbiased non-negative estimator $\pi(x)\, W$ of the target density $\pi(x)$, i.e. $\mathrm{E}(W) = 1$ for any $x \in \mathsf{X}$. For any integer $r \geq 1$, one may use an average of $r$ such estimators to construct the unbiased estimator $\pi(x)\,(W_1 + \ldots + W_r)/r$. In what follows, we assume $\underline{W} = (W_1, \ldots, W_r) \in \mathsf{W}^r$ is exchangeable with joint density $q_x(\underline{w})$. This accommodates the scenario where $W_1, \ldots, W_r$ are independent and distributed according to $q_x(w)$ so $q_x(\underline{w}) = q_x(w_1) \cdots q_x(w_r)$. We denote the associated kernel, acceptance probability and marginal acceptance probability by $\mathsf{P}_r$, $\alpha_r(x, \underline{w}; x', \underline{w}')$ and $\alpha_r(x, \underline{w})$, respectively.

Corollary 31 of Andrieu & Vihola (2016) shows that, for two positive integers $s \leq m$, asymptotic variances associated with $\mathsf{P}_m$ are at most those associated with $\mathsf{P}_s$ for $L^2(\pi)$ functions of the $x$-coordinate only. Given this ordering, it is natural to ask whether the decrease in asymptotic variance is sufficient to justify the extra computational expense of $\mathsf{P}_m$. Andrieu & Vihola (2015) show that the asymptotic variance of a pseudo-marginal algorithm ergodic average is bounded below by that of the algorithm in which $\pi$ is evaluated exactly; consequently, there must eventually be diminishing returns for any increase in $m$. For functions $\varphi \in L^2(\pi)$ of the $x$-coordinate only we are interested in $\mathrm{var}\{n^{-1}\sum_{k=1}^{n}\varphi(X_k)\} \approx n^{-1}\mathrm{V}(\varphi, \mathsf{P}_r)$. A reduction in variance equivalent to that obtained by increasing $r$ from $s$ to $m$ could instead be obtained by increasing $n$ by a factor of $\mathrm{V}(\varphi, \mathsf{P}_s)/\mathrm{V}(\varphi, \mathsf{P}_m)$. Since computational time is proportional to $n$, if it is also proportional to the number of samples per iteration, $r$, a natural way of comparing the two Markov kernels $\mathsf{P}_s$ and $\mathsf{P}_m$ is through their computational inefficiencies $s\mathrm{V}(\varphi, \mathsf{P}_s)$ and $m\mathrm{V}(\varphi, \mathsf{P}_m)$, respectively. As shown in the sequel, these quantities are not ordered in general but Theorem 1 below shows

that the quantities $r\{\mathrm{V}(\varphi, \mathsf{P}_r) + \mathrm{var}_\pi(\varphi)\}$ are. Since in many situations $\mathrm{var}_\pi(\varphi) \ll \mathrm{V}(\varphi, \mathsf{P}_r)$ this can be viewed as almost ordering computational inefficiencies.

THEOREM 1. *For positive integers $s \leq m$, the pseudo-marginal kernels $\mathsf{P}_s$ and $\mathsf{P}_m$ satisfy*

$$s\{\mathrm{V}(\varphi, \mathsf{P}_s) + \mathrm{var}_\pi(\varphi)\} \;\leq\; m\{\mathrm{V}(\varphi, \mathsf{P}_m) + \mathrm{var}_\pi(\varphi)\}, \tag{2}$$

*for any function $\varphi \in L^2(\pi)$ of the $x$-coordinate only.*

*Remark* 1. The inequality (A6) in the proof also implies through fairly simple manipulations that the average acceptance rates satisfy $\alpha_m \leq (m/s)\alpha_s$.

One interesting consequence is that the class of $L^2(\pi)$ functions with finite asymptotic variance, which is often not all of $L^2(\pi)$ (Lee & Łatuszyński, 2014), cannot be enlarged by increasing $m$.

COROLLARY 1. *In combination with Corollary 31 of Andrieu & Vihola (2016) we obtain that for $\varphi \in L^2(\pi)$, $\mathrm{V}(\varphi, \mathsf{P}_m) < \infty \iff \mathrm{V}(\varphi, \mathsf{P}_1) < \infty$.*

For a positive $\pi$-reversible Markov kernel $P$, $\mathrm{V}(\varphi, P) \geq \mathrm{var}_\pi(\varphi)$ for all $\varphi \in L^2(\pi)$. Consequently, Theorem 1 leads to the following generalisation of Proposition 4 of Bornn et al. (2017).

COROLLARY 2. *Let $s \leq m$ be positive integers and Markov kernel $\mathsf{P}_m$ be positive. For any function $\varphi \in L^2(\pi)$ of the $x$-coordinate only,* $\mathrm{V}(\varphi, \mathsf{P}_s) \leq (2m/s - 1)\,\mathrm{V}(\varphi, \mathsf{P}_m)$.

Positivity of some random-walk-based kernels can be verified via results of Doucet et al. (2015) and Sherlock (2016), which build upon Baxendale (2005). Independent Metropolis–Hastings pseudo-marginal kernels are always positive, as they are themselves independent Metropolis–Hastings kernels (Andrieu & Vihola, 2015).

### 2·4. *Tightness of the result*

The following proposition shows that the inequality in Theorem 1 cannot be improved in general, and that even if we consider averages of independent estimators the conjecture that $s\,\mathrm{V}(\varphi, \mathsf{P}_s) \leq m\,\mathrm{V}(\varphi, \mathsf{P}_m)$ is not true in general.

PROPOSITION 1. *There exist pseudo-marginal kernels and $\varphi \in L^2(\pi)$ such that*

1. *With negatively correlated $\underline{W}$,* $\mathrm{V}(\varphi, P_1) + \mathrm{var}_\pi(\varphi) = 2\{\mathrm{V}(\varphi, P_2) + \mathrm{var}_\pi(\varphi)\}$.
2. *With independent $\underline{W}$,* $\mathrm{V}(\varphi, P_1) > 2\mathrm{V}(\varphi, P_2)$.

The conjectured inequality, however, does hold in continuous time. Let $r \geq 1$ be an integer. We define the continuous-time Markov chain, with kernel $\widetilde{\mathsf{P}}_r$, as the Markov chain whose transitions are identical to those of the discrete-time kernel $\mathsf{P}_r$ but take place on a Poisson clock with unit rate. That is, if $\widetilde{X}_r(t)$ is the $x$-process of a continuous-time Markov chain with transition $\widetilde{\mathsf{P}}_r$ and $X_r(k)$ is the $x$-process of the discrete-time Markov chain with kernel $\mathsf{P}_r$, then $\widetilde{X}_r(t) = X_r(\mathrm{PP}[t])$ where $\{\mathrm{PP}[t]\}_{t\geq 0}$ designates a Poisson process with unit rate. For $\varphi \in L^2(\pi)$, the continuous-time asymptotic variance is defined as

$$\widetilde{\mathrm{V}}(\varphi, \mathsf{P}_r) = \lim_{T\to\infty} \frac{1}{T}\,\mathrm{var}\left[\int_0^T \varphi\Big\{\widetilde{X}_r(t)\Big\}\,dt\right].$$

PROPOSITION 2. *For positive integers $s \leq m$, the continuous-time chains satisfy*

$$s\widetilde{\mathrm{V}}(\varphi, \mathsf{P}_s) \;\leq\; m\widetilde{\mathrm{V}}(\varphi, \mathsf{P}_m), \qquad \varphi \in L^2(\pi).$$

Table 1: Computational efficiency ($\text{ESS}^* \times 10^3$): 90% confidence intervals

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| CE | $(7.7, 9.0)$ | $(5.7, 6.8)$ | $(4.9, 5.5)$ | $(4.1, 4.6)$ | $(3.6, 4.1)$ | $(3.2, 3.5)$ | $(2.9, 3.2)$ | $(2.6, 2.8)$ | $(2.4, 2.6)$ | $(2.1, 2.4)$ |

## 3. Numerical Studies

### 3·1. *Preliminaries*

We present in this Section two numerical studies. Several choices of proposal distributions are investigated and the situation when the computational time necessary to generate $m$ samples is not proportional to $m$, due to non-negligible computational overhead, is carefully examined. Computational efficiency is measured in terms of Effective Sample Size per unit of computational time; if the computational time to generate $m$ samples in each of $n$ iterations is exactly proportional to $nm$, then we define $\text{ESS}(\varphi, P) = n\text{var}_\pi(\varphi)/\{\text{V}(\varphi, P)\}$ and $\text{ESS}^*(\varphi, P) = \text{ESS}(\varphi, P)/(nm)$. Since $\text{ESS}(\varphi, P)$ and $\text{ESS}^*(\varphi, P)$ are intractable in general, we consistently estimate them below using realisations of the Markov chain with kernel $P$.

### 3·2. *Inverse Stochastic Heat Equation*

Let $x(t, u)$ designate the temperature at time $t \in [0, T]$ at the spatial location $u \in (0, 1)$. We consider the problem of reconstructing the initial temperature field $x(0, u)$ for $u \in (0, 1)$ from $N \geq 1$ noisy measurements at time $t = T$ distributed as $y_i = x(T, u_i) + \xi_i$ for some locations $\{u_i\}_{i=1}^N \subset (0, 1)$ and independent centred Gaussian samples $\{\xi_i\}_{i=1}^N$ with variance $\sigma_\xi^2$. The temperature field evolves according to the stochastic Heat equation

$$\partial_t x(t, u) = \Delta x(t, u) + \sigma \dot{W}(t, u) \tag{3}$$

with Dirichlet boundary $x(t, 0) = x(t, 1) = 0$; the process $\dot{W}$ is a space-time white noise (e.g. Hairer, 2009; arXiv:0907.4178). *A* priori, we use a truncated Karhunen–Loève expansion to model $x(0, \cdot)$, i.e. $x(0, u) = \sum_{k=1}^K \zeta_k \sin(k\,\pi\,u)$, where $\zeta_k \sim \mathcal{N}(0, k^{-2})$ are independent. Simulations were carried out by noting that, after finite discretisation in space, the evolution Equation (3) can be diagonalized and solved in the Fourier domain. The likelihood of a given initial temperature field can be unbiasedly evaluated by simulating $m$ trajectories and averaging the conditional likelihoods. The pseudo-marginal algorithms are started in a region of high posterior mass; we used a Crank–Nicholson proposal of the type $x_\star = \alpha\, x + (1 - \alpha^2)^{1/2}\, \zeta$, where $\zeta$ is distributed according to the prior distribution, with a value of $\alpha \in (-1, 1)$ chosen such that the acceptance rate when $m \gg 1$ is around $1/2$. The computational efficiency is taken as the minimum computational efficiency associated with 9 functions $x \mapsto x(0, i/10)$ for $i \in \{1, \ldots, 9\}$. Table 1 shows approximate 90% confidence intervals and, as expected, the computational efficiency is maximized for $m = 1$.

### 3·3. *Logistic regression using a latent Gaussian process*

We consider a logistic regression model with fixed effects and a latent Gaussian process, following exactly the approach of Sherlock (2016). The likelihood function is approximated by importance sampling with a data-dependent proposal distribution, similar to Filippone & Girolami (2014) and Giorgi et al. (2015). The parameter space has dimension 6 and the latent Gaussian process is required at $L = 144$ observation points. Whatever the value of $m$, at each iteration, creation of the importance sampling proposal involves a single $\mathcal{O}(L^3)$ Cholesky decomposition of an $L \times L$ matrix; each importance sample then costs $\mathcal{O}(L^2)$. For small values of $m$ the start-up cost dominates the cost of simulating $m$ importance samples.

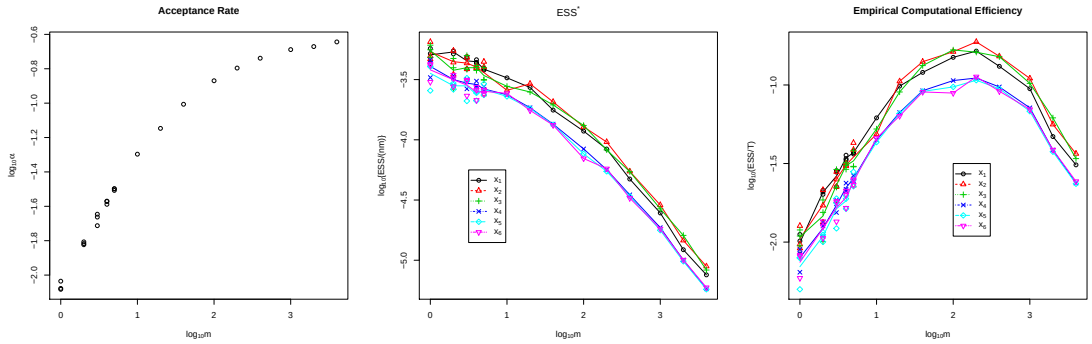


Fig. 1: Log quantities for each run against $\log_{10} m$ for (left) acceptance rate, (centre) hypothetical computational efficiency (ESS/$\{nm\}$) and (right) empirical computational efficiency (ESS/processing time). Each line in the latter two graphs corresponds to one of the six parameters.

The posterior mean and covariance matrix of the parameters were estimated from a trial run. For the random-walk pseudo-marginal Metropolis algorithm with $m = 100$, an approximately optimal scaling of $\lambda = 0.9$ was found. This scaling was then used in all runs recorded since it should also be approximately optimal for other values of $m$ (Sherlock, 2016), and thus help to control the Monte Carlo variability in empirical effective sample sizes. A set $\mathcal{M} \equiv \{1, 2, 3, 4, 5, 10, 20, 40, 100, 200, 400, 1000, 2000, 4000\}$ of numbers of importance samples was considered. Run lengths are given in the Supplementary Material and were chosen so as to keep the CPU time for each run between $10^5$ and $3 \times 10^5$ seconds and the effective sample sizes above $500$. Despite the large run lengths, the Monte Carlo variability was non-negligible for $m \leq 5$ and so three independent runs were performed for each of these $m$ values.

Figure 1 reports the average acceptance rate, the hypothetical computational efficiencies, ESS$^*$, and the empirical computational efficiencies, ESS divided by CPU time, for all six parameters. Due to the non-negligible computational overhead, the two efficiency measures are different. A pseudo-marginal independence sampler was also run and gave similar results. As the theory suggests, increasing from $s$ to $m$ samples never increases the acceptance rate by more than $m/s$ and the hypothetical computational efficiency is maximized at $m = 1$. However, due to the considerable start-up cost at each iteration, the empirical computational efficiency is maximized at around $m = 200$. Interestingly, it is at $m = 200$ that the cost of creating $m$ samples approximately matches the start-up cost.

## 4. AVERAGING VERSUS PARTICLE FILTERING

We have shown that if the computational cost of obtaining $m$ estimators is proportional to $m$ then it is close to optimal to choose $m = 1$ when averaging, at least when the asymptotic variance is finite. This is very different to Sherlock et al. (2015) and Doucet et al. (2015), who found under specific assumptions that when the likelihood function of a large number of observations is estimated via a particle filter, $m$ should be chosen so that the variance of the the log-likelihood estimator is controlled: the optimal choice of $m$ is consequently typically greater than one.

This fundamental difference arises because an estimator obtained using a particle filter is not an average, but a product of $T$ dependent averages of $m$ random variables. The relative variance

of an importance sampling estimator with $m$ samples is $C/m$ for some $C > 0$, whereas the relative variance of a particle filter estimate of the likelihood is of the form (Cérou et al., 2011),

$$\sum_{r=1}^{T} \left(\frac{1}{m}\right)^r \left(1 - \frac{1}{m}\right)^{T-r} C_r,$$

where $C_1, \ldots, C_T$ is is a non-negative sequence that often increases exponentially. It follows that increasing $m$ when $m$ is small can dramatically reduce the contributions of $C_2, \ldots, C_T$, even though by considering $m$ very large with $T$ fixed, the relative variance is $\mathcal{O}(C_1/m)$.

## SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of Propositions 1 and 2, and further details of the numerical studies.

## APPENDIX

If $P$ is reversible with respect to $\pi$, it can also be regarded as a self-adjoint operator on $L^2(\pi)$; the Dirichlet form associated with $P$ is defined as

$$\mathcal{E}_P(\varphi) \equiv \frac{1}{2} \int \pi(\mathrm{d}x) P(x, \mathrm{d}x') \{\varphi(x') - \varphi(x)\}^2.$$

*Proof of Theorem* 1. Recall Lemma 33 of Andrieu et al. (2016): if for two $\mu$-reversible Markov kernels $\Pi_1$ and $\Pi_2$ there exists $\varrho > 0$ satisfying $\mathcal{E}_{\Pi_2}(\varphi) \geq \varrho\, \mathcal{E}_{\Pi_1}(\varphi)$ for all $\varphi \in L^2(\mu)$, then

$$\varrho\{\mathrm{V}(\varphi, \Pi_2) + \mathrm{var}_\mu(\varphi)\} \;\leq\; \mathrm{V}(\varphi, \Pi_1) + \mathrm{var}_\mu(\varphi), \qquad \varphi \in L^2(\mu). \tag{A1}$$

To exploit this result, we construct two $\overline{\pi}$-reversible Markov kernels $\overline{P}_s$ and $\overline{P}_m$ on the extended space $\mathsf{X} \times \mathsf{W}^m \times [m]$ such that the $x$-marginal of the Markov chain with transition $\overline{P}_s$ (resp. $\overline{P}_m$) has the same law as the $x$-marginal of the Markov chain with transition $\mathsf{P}_s$ (resp. $\mathsf{P}_m$). By Lemma 33 of Andrieu et al. (2016), Theorem 1 follows once it is proved that for any $\varphi \in L^2(\overline{\pi})$ it holds that

$$\mathcal{E}_{\overline{P}_m}(\varphi) \;\leq\; \frac{m}{s}\, \mathcal{E}_{\overline{P}_s}(\varphi). \tag{A2}$$

We define the distribution $\overline{\pi}$, which depends on $s$ and $m$, through its density

$$\overline{\pi}(x, \underline{w}, k) \equiv \frac{1}{m}\, \pi(x) \left\{ \frac{w_k + \ldots + w_{k+s-1}}{s} \right\} q_x(\underline{w}) = \frac{1}{m}\, \pi(x)\, q_x(\underline{w})\, \mathrm{A}(\underline{w}, k) \tag{A3}$$

for $(x, \underline{w}, k) \in \mathsf{X} \times \mathsf{W}^m \times [m]$; the indices in (A3) and henceforth are to be understood modulo $m$, and we have used the notation $\mathrm{A}(\underline{w}, k) = (w_k + \ldots + w_{k+s-1})/s$. The Metropolis–Hastings kernel $\overline{P}_s$ proposes a move $(x, \underline{w}, k) \mapsto (X', \underline{W}', K')$ by first generating $(X', \underline{W}') \sim q(x, dx')\, q_{x'}(d\underline{w}')$ and then choosing $K'$ uniformly at random in $[m]$, i.e. the proposal density is

$$\overline{q}_s(x, \underline{w}, k; x', \underline{w}', k') \equiv q(x, x')\, q_{x'}(\underline{w}')\, (1/m).$$

The proposed $(X', \underline{W}', K')$ is accepted with the usual Metropolis–Hastings probability

$$\overline{\alpha}_s(x, \underline{w}, k; x', \underline{w}', k') \;=\; 1 \wedge \left\{ r(x, x') \, \frac{w'_{k'} + \ldots + w'_{k'+s-1}}{w_k + \ldots + w_{k+s-1}} \right\}. \tag{A4}$$

The Metropolis–Hastings kernel $\overline{P}_m$ differs from $\overline{P}_s$ in the way $K'$ is proposed. It proposes a move $(x, \underline{w}, k) \mapsto (X', \underline{W}', K')$ by first generating $(X', \mathsf{W}') \sim q(x, dx')\, q_{x'}(d\underline{w}')$ and then choosing $K' \in [m]$ such that $\mathrm{pr}\,(K' = k') \propto A(\underline{w}', k')$. Since for any $\underline{w} \in \mathsf{W}^m$ we have $\mathrm{A}(\underline{w}, 1) + \ldots + \mathrm{A}(\underline{w}, m) = w_1 +$

$\ldots + w_m$, the proposal density is

$$\overline{q}_m(x, \underline{w}, k; x', \underline{w}', k') \equiv q(x, x')\, q_{x'}(\underline{w}')\, \frac{\mathrm{A}(\underline{w}', k')}{w_1' + \ldots + w_m'}.$$

The proposed $(X', \underline{W}', K')$ is accepted with the usual Metropolis–Hastings probability

$$\overline{\alpha}_m(x, \underline{w}, k; x', \underline{w}', k') = 1 \wedge \left\{ r(x, x')\, \frac{w_1' + \ldots + w_m'}{w_1 + \ldots + w_m} \right\}. \tag{A5}$$

From Equations (A4)–(A5) it follows that the $x$-coordinates of the Markov chains with transition $\overline{P}_s$ and $\overline{P}_m$ equal in law, respectively, the $x$-coordinates of the Markov chains with transitions $P_s$ and $P_m$. To conclude the proof, we now prove inequality (A2); it suffices to prove that $\overline{P}_m(x, \underline{w}, k; \mathrm{d}x', \mathrm{d}\underline{w}', k')$ is at most $(m/s)\, \overline{P}_s(x, \underline{w}, k; \mathrm{d}x', \mathrm{d}\underline{w}', k')$ for any $(x, \underline{w}, k) \neq (x', \underline{w}', k')$, i.e.

$$\frac{\mathrm{A}(\underline{w}', k')}{w_1' + \ldots + w_m'}\, \overline{\alpha}_m(x, \underline{w}, k; x', \underline{w}', k') \leq (m/s)\, \{m^{-1}\, \overline{\alpha}_s(x, \underline{w}, k; x', \underline{w}', k')\}. \tag{A6}$$

From (A4)–(A5), this is equivalent to showing that $1 \wedge \{r(x, x')\, (w_1' + \ldots + w_m')/(w_1 + \ldots + w_m)\}$ is at most

$$\left\{ \frac{w_1' + \ldots + w_m'}{s\, A(\underline{w}', k')} \right\} \wedge \left[ r(x, x')\, \frac{w_1' + \ldots + w_m'}{w_1 + \ldots + w_m} \left\{ \frac{w_1 + \ldots + w_m}{s\, A(\underline{w}, k)} \right\} \right],$$

and since the two quantities inside curly brackets are at least one, the conclusion follows. □

## REFERENCES


ANDRIEU, C., DOUCET, A. & HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Statist. Soc. B* **72**, 269–342.

ANDRIEU, C., LEE, A. & VIHOLA, M. (2016). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli* , to appear.

ANDRIEU, C. & ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**, 697–725.

ANDRIEU, C. & VIHOLA, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.* **25**, 1030–1077.

ANDRIEU, C. & VIHOLA, M. (2016). Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.* **26**, 2661–2696.

BAXENDALE, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.* **15**, 700–738.

BEAUMONT, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160.

BORNN, L., PILLAI, N. S., SMITH, A. & WOODARD, D. (2017). The use of a single pseudo-sample in approximate Bayesian computation. *Statistics and Computing* **27**, 583–590.

CÉROU, F., DEL MORAL, P. & GUYADER, A. (2011). A nonasymptotic theorem for unnormalized Feynman–Kac particle models. *Ann. Inst. H. Poincaré Probab. Statist.* **47**, 629–649.

DOUCET, A., PITT, M. K., DELIGIANNIDIS, G. & KOHN, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102**, 295–313.

FILIPPONE, M. & GIROLAMI, M. (2014). Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 2214–2226.

GIORGI, E., SESAY, S. S. S., TERLOUW, D. J. & DIGGLE, P. J. (2015). Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *J. R. Statist. Soc. A* **178**, 445–464.

LEE, A. & ŁATUSZYŃSKI, K. (2014). Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika* **101**, 655–671.

SHERLOCK, C. (2016). Optimal scaling for the pseudo-marginal random walk Metropolis: insensitivity to the noise generating mechanism. *Methodol. Comput. Appl. Probab.* **18**, 869–884.

SHERLOCK, C., THIERY, A., ROBERTS, G. O. & ROSENTHAL, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43**, 238–275.